AF&PA®
AMERICAN FOREST & PAPER ASSOCIATION
GROWING WITH AMERICA SINCE 1861

January 11, 2002

Dr. Michael Shelby
Director CERHR
National Institute of Environmental Health Sciences
79 T.W. Alexander Dr.
Bldg. 4401 RM 103
PO Box 12233, MD EC-32
Research Triangle Park, NC 27709

JAN 1 6 2002

Dear Dr. Shelby:

At the October 2001, meeting of the CERHR Methanol Expert Panel, Dr. David Hoel presented comments on statistical issues relating to the HEI-sponsored primate study conducted by Dr. Thomas Burbacher. Based on his analysis, Dr. Hoel concluded that the few scattered positive statistical findings were what one would expect to occur simply by chance.

During the panel's discussions, a number of members also raised questions about the statistical analyses performed in the primate study. The panel's draft report suggested that a more rigorous analysis of the data would be helpful in evaluating the evidence of whether or not methanol is a developmental toxicant in monkeys.

At the request of AF&PA, Dr. Hoel has performed further detailed statistical analysis of the Burbacher data. We want to share this analysis with you and members of the panel because we believe it may help fill the important information need identified by the panel.

Dr. Hoel has raised four principal concerns with the Burbacher et al. statistical analyses. First, Burbacher et al. failed to adjust properly for the numerous statistical tests that were undertaken (the problem of multiple comparisons). This has the effect of markedly elevating the false positive error rate above its nominal 5% value. Second, Burbacher et al. employed analysis of variance (ANOVA) followed by multiple pairwise t-tests, a methodology that failed to account properly for the graded dose design of the experiment (i.e., animals were exposed to either 0, 200, 600, or 1,800 ppm methanol).

Third, Burbacher et al. simply assumed that all of the data were normally distributed with equal variances across the exposure groups. When this assumption is not valid, it can lead to erroneous conclusions regarding the presence of exposure-related effects. Finally, Burbacher et al. conducted many more *post-hoc* tests on subsets of animals (e.g., by cohort), without apparent concern for the extremely small sample sizes this produced (e.g., as small as 2 or 3 animals per group), and with selective reporting of results, thus concealing the true extent of the additional analyses that were undertaken. This further exacerbates the already severe multiple comparison problem.

1111 Nineteenth Street, NW, Suite 800 ▪ Washington, DC 20036 ▪ 202 463-2700 Fax: 202 463-2785 ▪ www.afandpa.org

In the attached report, Dr. Hoel has summarized his reanalysis of the Burbacher et al. study data utilizing statistical methodology that is optimally matched to the study's experimental design, with appropriate adjustment of the false positive error rate for the multiple comparisons problem. He has also utilized non-parametric statistical methodology that is robust to departures from normality and equality of variances to ensure that such departures will not invalidate any conclusions. Finally, he has noted those cases in the *post-hoc* testing reported by Burbacher et al. in which the sample sizes were so small as to call the legitimacy of virtually any statistical analysis into question.

The results of Dr. Hoel's reanalysis are clear-cut and consistent. There were no endpoints for which statistically significant findings were observed. Indeed, his overall conclusion is that "Burbacher et al. primate study showed no reproductive or offspring developmental effects of methanol exposure."

I hope Dr. Hoel's analysis will be helpful to the panel. If you have any questions, please contact me at 202-463-2587.

Thanks you

Sincerely

John L. Festa

John L. Festa
Senior Scientist, Ph.D.

Statistical Analysis of Reported Developmental Effects in the HEI Report Entitled:

Reproductive and Offspring Developmental Effects Following Maternal Inhalation Exposures to Methanol in Nonhuman Primates

by Thomas Burbacher et al.

Prepared by

David G. Hoel, Ph.D.

30 December 2001

Prepared on behalf of the American Forest and Paper Association

Statistical Analysis of the Burbacher et al. (1999) Methanol Primate Study
David G. Hoel, Ph.D.
December 30, 2001

Background:
The statistical analysis of the Burbacher et al. study (Reproductive and Offspring Developmental Effects following Maternal Inhalation Exposure to Methanol in Nonhuman Primates by Thomas Burbacher et al., Health Effects Institute Report Number 89, October 1999) has major problems. These problems are such that an erroneous picture is given that there may be an association between behavioral development and maternal exposure to methanol in primates. The statistical problems can be described as the following four issues:

1) There were multiple statistical tests carried out without any adjustment for multiple comparisons, resulting in markedly inflated false positive error rates.

2) For a given outcome, both ANOVA and multiple t-tests were performed on the data. However, the study involved four experimental groups receiving graded doses of methanol (0 ppm, 200 ppm, 600 ppm and 1800 ppm), but the statistical methods that were used did not incorporate this design information.

3) The data were assumed to be normally distributed with the same variance across the experimental groups.

4) Post-hoc tests were carried out on subsets of animals (e.g., by sex or by cohort) without apparent concern for small sample sizes (e.g., as small as 2 or 3 per group) and the large number of additional tests (about 500), again without any adjustment for multiple testing. Only supposed positive test results were reported, which concealed the extent of the additional analyses.

A summary of the statistical testing in the Burbacher study is given in the following two tables.

Table A
Summary of Statistical Testing
Part II: Developmental Effects, Primary Analyses

	Number of Experiments	Statistical Tests	Expected Positives	Observed Positives
Physical Measures				
ANOVA plus 4 contrasts	12	60	3	0
Behavioral Measures				
ANOVA plus 4 contrasts	14	70	3.5	4*
4 contrasts only	5	20	1	0
Total	31	150	7.5	4

* A9 (c vs. 600 ppm & c vs. total exposed), A13 (c vs. 1800 ppm) and A26 (c vs. 600 ppm).

Note: linear dose-response testing was reported only for A13 (p=0.04) and A15 (p=0.08)

Table B

Specific Statistically Significant Results:
Primary and Secondary Analyses

Experiment	ANOVA	Sum*	200 ppm	600 ppm	1800 ppm	Linear	Secondary**
A9[a]	n.s.	0.03	n.s.	0.01	n.s.		none
A13[b]	n.s.	n.s.	n.s.	n.s.	0.04	0.04	m/f [1]
A15[b]	n.s.	n.s.	n.s.	n.s.	n.s.	0.08	m/f [2]
A17[c]	n.s.	n.s.	n.s.	n.s.	n.s.		m/f [3]
A26[d]	n.s.	n.s.	n.s.	0.03	n.s.		cohort [4]

* Sum is control group contrasted with the combined treated groups.
** Secondary Analyses means a new analysis incorporating additional factors such as gender is reported, which, presumably, was motivated by the results of the original analysis.

[a] Neonatal behavioral scale: Behavior state factor
[b] Visually directed reaching: vs. age (A13), vs gestational length (A15)
[c] Observations of motor milestones in playroom
[d] Recognition memory assessment
[1] Males: Control vs. 600 ppm p=0.007, Control vs. 1800 ppm p=0.03; Females: n.s.
[2] Males: Control vs. 600 ppm p=0.04, Control vs. 1800 ppm p=0.04; Females: n.s.
[3] Males: Control vs. 600 ppm p=0.02; Female: Control vs. 200 ppm p=0.01, Control vs. 600 ppm p=0.004, Control vs. Total exposed p=0.008.
[4] Cohort 1: n.s.; Cohort 2: Control vs. 200 ppm p=0.04, Control vs. 600 ppm p=0.0001, Control vs. 1800 ppm p=0.03, Control vs. Total Exposed p=0.002.
Note: n.s. => p > 0.05

Statistical Methods:
The experimental data should be analyzed as follows. For each outcome variable the data should be statistically tested using a single test of the null hypothesis of no experimental effect of methanol versus either a linear trend with exposure or an ordered alternative (i.e., a monotonic dose-response). Tests which do this are more powerful than individual repeated pairwise t-tests with p-values properly corrected for multiple testing. Quoting Williams (Biometrics 27:104,1971).

'These (i.e., t-tests) are less powerful than tests which take some account of the dependence of response on dose. Moreover they sometimes lead to difficulties in interpretation. For example, what is the experimenter to make of differences from control which are significant at some dose level but not at one or more higher doses? Although he may not be prepared to assume a regression model for the dependence of response on dose, he certainly expects the response to be compatible with this expectation.

Tests using an ordered alternative represent the appropriate middle ground between ANOVA with multiple t-tests and dose-response modeling with a specific model such as a linear dose-response relationship. Ordered alternative testing can be done either by assuming normally distributed data with equal variances across dose groups or by using nonparametric test procedures. These two approaches seem to give similar results as long as the assumptions of normality and equal variances in the dose groups are not violated.

For normally distributed data an ordered alternative test is given by Williams (Biometrics 28:519-531,1972). A test statistic is calculated for each non-control dose group. Beginning with the test statistic at the highest dose (e.g., 1800 ppm) a comparison is made with a tabulated critical value. If the test statistic exceeds the appropriate critical value, one moves to the test statistic for the next largest dose group. This procedure is repeated until the test statistic no longer exceeds its corresponding critical value. The dose corresponding to the first test statistic that does not exceed its critical value becomes the no effect level. If none of the test statistics exceed their critical values, then the null hypothesis of no dose-related effect is accepted.

For nonparametric testing we have the statistical tests of Jonckheere (Biometrika 41:133-145, 1954) and Shirley (Biometrics 33:386-389, 1977). For linear trend testing, a nonparametric alternative to linear regression is the nonparametric trend test given by Cuzick (Stat in Med 4:87-90 1985). As described above for the Williams test, the Shirley nonparametric test also provides an estimate of the lowest dose group for which there is a significant increase (or decrease) from the controls. The Jonckheere test is the more common of these nonparametric tests. It, as well as the Shirley and Cuzick tests, has the advantage of not requiring the assumption that the data are normally distributed. Furthermore the test is based on a statistic which is itself normally distributed under the null hypothesis which means that specific tabulations of critical test values are not required.

From Table A we see that 31 experiments were carried out and analyzed. Therefore in order to obtain an experiment-wide false positive error rate of $p = 0.05$ we should set the p-value for each experiment at $p = 0.05/31 = 0.002$. Essentially the same individual p-value of 0.002 would apply if we only considered the 19 behavioral tests. Table B summarizes the results from the original ANOVA and multiple t-test analyses. We will apply the above mentioned statistical procedures to the data from those experiments in which the authors had suggested there may be a methanol effect.

Results:

1) Experiment A9: Neonatal Behavior Scale: Behavioral State

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	1.188	2.038	2.097
critical value (p=0.05)	2.042	2.106	2.131
critical value (p=0.01)	2.750	2.792	2.809
critical value (p=0.002)	3.385	N/A	N/A

Note: if the test statistic value exceeds its critical value then the result is statistically significant (e.g., if 2.097 were > 2.131 we would have significance at the 0.05 level for the 1800 ppm dose group.) Also the critical values which are tabulated (see Williams' papers) are not available for the very small p-value appropriate for the Burbacher et al. study, i.e., $p = 0.002$. Only the critical value for the lowest dose group (i.e. 3.385) is obtainable from a t-table while the others require more complicated numerical integrations which are not undertaken here. However, if the calculated test statistics are not significant at the p = 0.01 level, then it is obvious that they must also fail to achieve statistical significance at the p = 0.002 level.

The result of the Williams test is that there is no methanol effect at p=0.05 and certainly not at the appropriate p value of 0.002.

Non-parametric results:

Jonckheere test of ordered alternatives: test statistic Z=1.8634 and p=0.06
Cuzick's non-parametric linear trend test: test statistic Z=1.80 and p=0.07

We conclude that for this endpoint there is no methanol effect.

2) Experiment A13: Visually Directed Reaching: (Using actual age)

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	0.515	0.621	2.100
critical value (p=0.05)	2.042	2.106	2.131
critical value (p=0.01)	2.750	2.792	2.809

The result of the Williams test is that there is no methanol effect at p=0.05 and certainly not at the appropriate p value of 0.002.

Non-parametric results:

Jonckheere test of ordered alternatives: test statistic Z=1.433 and p=0.15
Cuzick's non-parametric linear trend test: test statistic Z=1.46 and p=0.14

We conclude that for this endpoint there is no methanol effect.

Post-hoc analysis using only male primates:

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	1.049	1.949	1.743
critical value (p=0.05)	2.262	2.351	2.364
critical value (p=0.01)	3.250	3.329	3.334

The result of the Williams test is that there is no methanol effect at p=0.05 and certainly not at the appropriate p value of 0.002.

Non-parametric results:

Jonckheere test of ordered alternatives: test statistic Z=1.992 and p=0.05
Cuzick's non-parametric linear trend test: test statistic Z=1.84 and p=0.07

It should be noted that there are so few animals in the male primate subset (3, 5, 3, 2 animals in the 4 dose groups), that there is a serious question concerning the validity of any statistical testing.

We conclude that for this endpoint there is no methanol effect among male primates.

2a) Experiment A15: Visually Directed Reaching: (Using age since conception)

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	-1.21	- 0.610	0.593
critical value (p=0.05)	2.042	2.106	2.131
critical value (p=0.01)	2.750	2.792	2.809

The result of the Williams test is that there is no methanol effect at p=0.05 and certainly not at the appropriate p value of 0.002.

Non-parametric results:
Jonckheere test of ordered alternatives: test statistic Z=0.092 and p=0.93
Cuzick's non-parametric linear trend test: test statistic Z=0.05 and p=0.96

We conclude that for this endpoint there is no methanol effect.

Post-hoc analysis using only male primates:

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	0.043	0.771	1.327
critical value (p=0.05)	2.262	2.351	2.364
critical value (p=0.01)	3.250	3.329	3.334

The result of the Williams test is that there is no methanol effect at p=0.05 and certainly not at the appropriate p value of 0.002.

Non-parametric results:
Jonckheere test of ordered alternatives: test statistic Z=1.219 and p=0.22
Cuzick's non-parametric linear trend test: test statistic Z=1.22 and p=0.22

It should be noted that there are so few animals in the male study (3,5,3,2 animals in the 4 dose groups) that there is a serious question concerning the validity of any statistical testing.

We conclude that for this endpoint there is no methanol effect among male primates.

3) Experiment A17: Motor Milestones in Playroom

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	1.578	1.533	1.578
critical value (p=0.05)	2.042	2.106	2.131
critical value (p=0.01)	2.750	2.792	2.809

The result of the Williams test is that there is no methanol effect at p=0.05 and certainly not at the appropriate p value of 0.002.

Non-parametric results:
Jonckheere test of ordered alternatives: test statistic Z=-0.907 and p=0.36
Cuzick's non-parametric linear trend test: test statistic Z=-0.86 and p=0.39

We conclude that for this endpoint there is no methanol effect.

Post-hoc analysis using only female primates:

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	2.158	2.289	2.473
critical value (p=0.05)	2.080	2.181	2.212
critical value (p=0.01)	2.898	2.951	2.986

The result of the Williams test is that there is no methanol effect at p=0.01 and certainly not at the appropriate p value of 0.002.

Non-parametric results:
Jonckheere test of ordered alternatives: test statistic Z=-0.945 and p=0.345
Cuzick's non-parametric linear trend test: test statistic Z=-1.08 and p=0.28

It should be noted that there are so few animals in the female study (5, 4, 5, 7 animals in the 4 dose groups) that there is a question concerning the validity of any statistical testing and the normality assumption. (Note: The difference in the level of significance observed with the nonparametric and normal tests is likely due to the fact that the group standard deviations are not similar which violates the assumption of equal variances for the normal theory test. Therefore the nonparametric test is more reliable.)

We conclude that for this endpoint there is no methanol effect among female primates.

Post-hoc analysis using only male primates:

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	-0.129	0.657	0.587

critical value (p=0.05)	2.262	2.351	2.364
critical value (p=0.01)	3.250	3.329	3.334

The result of the Williams test is that there is no methanol effect at p=0.05 and certainly not at the appropriate p value of 0.002.

Non-parametric results:

Jonckheere test of ordered alternatives: test statistic Z=0.064 and p=0.95

Cuzick's non-parametric linear trend test: test statistic Z=-0.43 and p=0.67

It should be noted that there are so very few animals in the male study (3, 5, 3, 2 animals in the 4 dose groups) that there is a serious question concerning the validity of any statistical testing.

We conclude that for this endpoint there is no methanol effect among male primates.

4) Experiment A26: Recognition Memory Assessment

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	0.533	1.246	1.282
critical value (p=0.05)	2.042	2.106	2.131
critical value (p=0.01)	2.750	2.792	2.809

The result of the Williams test is that there is no methanol effect at p=0.05 and certainly not at the appropriate p value of 0.002.

Non-parametric results:

Jonckheere test of ordered alternatives: test statistic Z=-1.063 and p=0.29

Cuzick's non-parametric linear trend test: test statistic Z=-1.09 and p=0.28

We conclude that for this endpoint there is no methanol effect.

Post-hoc analysis using only cohort 2 primates:

Normal distribution assumption:

	Exposure Group		
	200 ppm	600 ppm	1800 ppm
test statistic value	0.989	2.663	2.663
critical value (p=0.05)	2.201	2.293	2.327
critical value (p=0.01)	3.106	3.186	3.211

The result of the Williams test is that there is no methanol effect at p=0.01 and certainly not at the appropriate p value of 0.002.

Non-parametric results:

Jonckheere test of ordered alternatives: test statistic Z=1.34 and p=0.18

Cuzick's non-parametric linear trend test: test statistic Z=1.47 and p=0.14

It should be noted that there are so few animals in the cohort 2 study (3, 4, 5, 4 animals in the 4 dose groups) that there is a question concerning the validity of any statistical testing and especially the normality assumption.

We conclude that for this endpoint there is no methanol effect among female primates.

5) Other endpoints:

The other experimental endpoints did not show a significant methanol effect except for the measure of 'duration of pregnancy'. The control group had one 'post-maturity' male primate that acted as an outlier and caused a significant difference between average pregnancy duration in the controls and the methanol exposed groups. Removal of this single outlier animal from the analysis resulted in there being no statistically significant differences in the duration of pregancy between the control and exposed animals.

Overall Conclusion:

The ordered alternative test procedures used here are more powerful than ANOVA followed by multiple pairwise t-tests. Nevertheless, the results are quite similar in that no experimental endpoint showed a significant effect of methanol exposure at the p=0.002 level using the ordered alternative statistical methods or ANOVA. The Cuzick nonparametric test for a linear trend was also carried out on this data set with the results essentially identical to those from the nonparametric Jonckheere test for ordered alternatives.

Using statistical procedures appropriately matched to the design of the Burbacher et al. experiment, and appropriate adjustments for multiple comparisons to control the experiment-wide false positive error rate, we see that there were no statistically significant developmental or behavioral effects at all among the many specific endpoint evaluations carried out on these primates. One can therefore only conclude that the Burbacher et al. primate study showed no reproductive or offspring developmental effects of methanol exposure.

References:

Cuzick, J. 1985. A Wilcoxon-type test for trend. *Statistics in Medicine* 4: 87-90.

Jonckheere, A.R. 1954. A distribution-free k-sample test against ordered alternatives. *Biometrika* 41:133-145.

Shirley, E. 1977. A non-parametric equivalent of William's test for contrasting increasing dose levels of a treatment. *Biometrics* 33:386-389.

Williams, D.A. 1971. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics* 27:103-117.

Williams, D.A. 1972. The comparison of several dose levels with a zero dose control. *Biometrics* 28:519-531.